OPINION

Why we need a centralized repository for isotopic data

Jonathan N. Pauli[a,1], Seth D. Newsome[b], Joseph A. Cook[c], Chris Harrod[d], Shawn A. Steffan[e,f], Christopher J. O. Baker[g], Merav Ben-David[h], David Bloom[i], Gabriel J. Bowen[j], Thure E. Cerling[j], Carla Cicero[k], Craig Cook[h], Michelle Dohm[l], Prarthana S. Dharampal[f], Gary Graves[m,n], Robert Gropp[o], Keith A. Hobson[p], Chris Jordan[q], Bruce MacFadden[r], Suzanne Pilaar Birch[s,t], Jorrit Poelen[u], Sujeevan Ratnasingham[v], Laura Russell[i], Craig A. Stricker[w], Mark D. Uhen[x], Christopher T. Yarnes[y], and Brian Hayden[z]

Stable isotopes encode and integrate the origin of matter; thus, their analysis offers tremendous potential to address questions across diverse scientific disciplines (1, 2). Indeed, the broad applicability of stable isotopes, coupled with advancements in high-throughput analysis, have created a scientific field that is growing exponentially, and generating data at a rate paralleling the explosive rise of DNA sequencing and genomics (3). Centralized data repositories, such as GenBank, have become increasingly important as a means for archiving information, and "Big Data" analytics of these resources are revolutionizing science and everyday life.

However, to date a centralized database for the management of isotopic data does not exist. We believe that the absence of such a resource has impeded research progress through the unnecessary duplication of effort, restricted the near-boundless application of stable isotopes, and curtailed the exchange of information among researchers. The creation of such a centralized database would be more than a silo for data; it would be a dynamic resource to unite disciplinary fields and answer pressing questions in agriculture, animal sciences, archaeology, anthropology, ecology, medicine, nutrition, physiology, paleontology, forensics, and earth and planetary sciences. We believe that a centralized database for isotopes would accelerate and enhance such global and multidisciplinary endeavors, thus broaden the reach of isotope science. Here, we—a group of stable isotope scientists, data managers, museum curators, journal editors, and educators—offer a vision for the public repository's identity, structure, and long-term sustainability.

The Need for IsoBank

Stable isotopes play a ubiquitous role in modern science; hence, the benefits of IsoBank are potentially immense. Isotopes have been used to construct isoscapes, continental or oceanic scale maps of isotope ratios in ground water and organic materials, transforming the fields of ecology and food and forensic science (4). Stable isotopes have a long history of use by archaeologists to reconstruct our past movements and diet and the rise and fall of civilizations (5), and by nutritionists to assess our current health (6). They are used by earth scientists to document the environmental and evolutionary history of the Earth, and by ecologists and physiologists to track the flux of nutrients between and within ecosystems (7) and individuals (8). More recently, researchers have begun to harness large isotopic datasets to address questions of global relevance—global nitrogen cycling (9) or continental climate variation (10).

Yet, the syntheses of isotope data across broad spatial and temporal scales and across disciplinary fields has generally been hampered by the difficulty of efficiently procuring large datasets from the published literature. This is compounded with the reality that most of the isotope data that currently exists are not, and may never be, published in peer-reviewed journals. Other relevant data are published in articles going back decades, but are effectively inaccessible to researchers. IsoBank would provide a route to enhance

[a]Department of Forest and Wildlife Ecology, University of Wisconsin–Madison, Madison, WI 53706; [b]Center for Stable Isotopes, Department of Biology, University of New Mexico, Albuquerque, NM 87131; [c]Museum of Southwestern Biology, Department of Biology, University of New Mexico, Albuquerque, NM 87131; [d]Instituto de Ciencias Naturales Alexander von Humboldt, Universidad de Antofagasta, Antofagasta 1270300, Chile; [e]US Department of Agriculture, Agricultural Research Service, Madison, WI 53706; [f]Department of Entomology, University of Wisconsin–Madison, Madison, WI 53706; [g]Department of Computer Science, University of New Brunswick, Saint John, NB, Canada E2L 4L5; [h]Department of Zoology and Physiology, University of Wyoming, WY 82071; [i]VertNet/iDigBio, Florida Museum of Natural History, University of Florida, Gainesville, FL 32611; [j]Department of Geology and Geophysics, University of Utah, Salt Lake City, UT 84112; [k]Museum of Vertebrate Zoology, University of California, Berkeley, CA 94720; [l]Public Library of Science, San Francisco, CA 94111; [m]Department of Vertebrate Zoology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20013-7012; [n]Center for Macroecology, Evolution, and Climate, Natural History Museum of Denmark, University of Copenhagen, DK-2100 Copenhagen, Denmark; [o]American Institute of Biological Sciences, Washington, DC 20005; [p]Environment Canada, Saskatoon, SK Canada S7N 3H5; [q]Texas Advanced Computing Center, The University of Texas at Austin, Austin, TX 78758; [r]Florida Museum of Natural History, University of Florida, Gainesville, FL 32611; [s]Department of Anthropology, University of Georgia, GA 30602; [t]Department of Geography, University of Georgia, GA 30602; [u]Private address, Oakland, CA 94610; [v]Centre for Biodiversity Genomics, University of Guelph, Guelph, ON, Canada N1G 2W1; [w]US Geological Survey, Fort Collins Science Center, Denver, CO 80225; [x]George Mason University, Fairfax, VA 22030; [y]Stable Isotope Facility, University of California, Davis, CA 95616; and [z]Biology Department, University of New Brunswick, Fredericton, NB, Canada E3B 5A3

The authors declare no conflict of interest.

Any opinions, findings, conclusions, or recommendations expressed in this work are those of the authors and have not been endorsed by the National Academy of Sciences.

[1]To whom correspondence should be addressed. Email: jnpauli@wisc.edu.

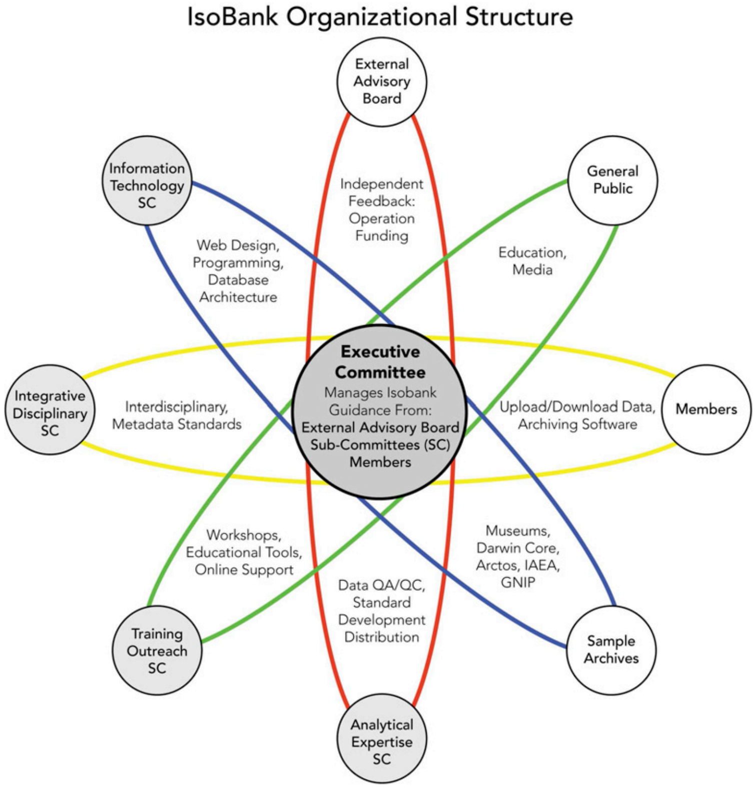

Fig. 1. Organizational structure for the proposed IsoBank. A central executive group would oversee four subcommittees (SC): Information technology, integrative disciplinary, education and training, and analytical expertise. GNIP, Global Network of Isotopes in Precipitation; IAEA, International Atomic Energy Association; QA/QC, quality assurance/quality control.

interdisciplinary research and a portal to published and unpublished datasets. Such a resource, then, could enhance our understanding of human history, our predictions of global change, the diagnoses and treatment of human disease, and the study of our planet and solar system.

We envisage IsoBank as both an aggregator and a repository of isotopic data. It should be an online, openly accessible database, with isotope measurements indexed via discipline-specific metadata. When possible, data deposited in IsoBank should be linked to archived samples and specimens. IsoBank will function as a universal resource, and allow scientists to verify, replicate, compare, extend, and integrate data across studies. In the same way that GenBank filled an immediate need within the field of genetics, IsoBank will consolidate and organize the broad and growing number of disciplines that have the potential to use stable isotope measurements. IsoBank should be networked internationally with core isotope laboratories, government-funded science agencies, and peer-reviewed journals to foster collaborations and ensure sustainability.

Organizational Structure

The structure of IsoBank requires the recognition of the breadth of research conducted with stable isotopes and the inclusion of a broad group of researchers, educators, museum curators, and data repository experts to develop and oversee its operation (Fig. 1). The efforts of this group would be targeted by a team of project coordinators, each heading one subcommittee (below) and overseen by an independent advisory board consisting of experienced isotope scientists and database mangers.

We envision at least four subcommittees (Fig. 1): (*i*) Information Technology: programmers, database architects, and web-designers who would build and maintain a high-capacity and user-friendly platform for IsoBank; (*ii*) Education and Training: specialists who would lead workshops, train potential users, provide online support, and promote professional development and outreach; (*iii*) Analytical Expertise: a consortium of core laboratories that analyze large volumes and diverse types of samples, to ensure rigorous data standards and enhance cohesion and communication among independent analytical facilities and thereby facilitate the development of disciplinary standards for data quality and laboratory operations, addressing the deficiency in isotopic investigations; and (*iv*) Integrative Disciplinary: leaders in relevant fields, presenting the diverse use of isotopes across disciplines, who would help craft rigorous metadata standards, identify disciplinary terminology, and reinforce its use.

Data Storage and Metadata Structure

For any repository to be useful, the data must be reliable, accessible—ideally in a machine-readable format—and have agreed-upon semantics for the data and metadata fields. A hierarchical design with relevant metadata fields would enable the alignment of isotope data from diverse research areas, and allow data to be traced back to analytical laboratories to facilitate independent quality assurance/quality control reviews.

Documenting the ontology of metadata will be one of the great challenges for IsoBank (e.g., ref. 11). Such a task is particularly challenging, given the broad range of disciplines involved and the importance of such metadata in statistical analyses (Fig. 2). Where possible, IsoBank should use existing ontologies, facilitating current and future integration with existing databases. For example, IsoBank could assign integrative taxonomic information system (https://www.ITIS.gov) serial numbers to organismal submissions that are then linked to a geographic distribution, evolutionary, or ecological relationships (12–14). We envision a database revolving around three primary informational subunits—user, sample, and analytical—each of which will be associated with core metadata terms, which could be further classified where required (Fig. 2).

Isobank would need to seamlessly incorporate user information. Similar to other data repositories, IsoBank users should be able to link existing online profiles, ideally ORCID (https://orcid.org), to their IsoBank profile and data submissions. Just as active data repositories, such as FigShare (https://figshare.com) and Dryad (www.datadryad.org), allocate a DOI for data loaded to their site, IsoBank should also allow users to receive recognition through DOI citations when data

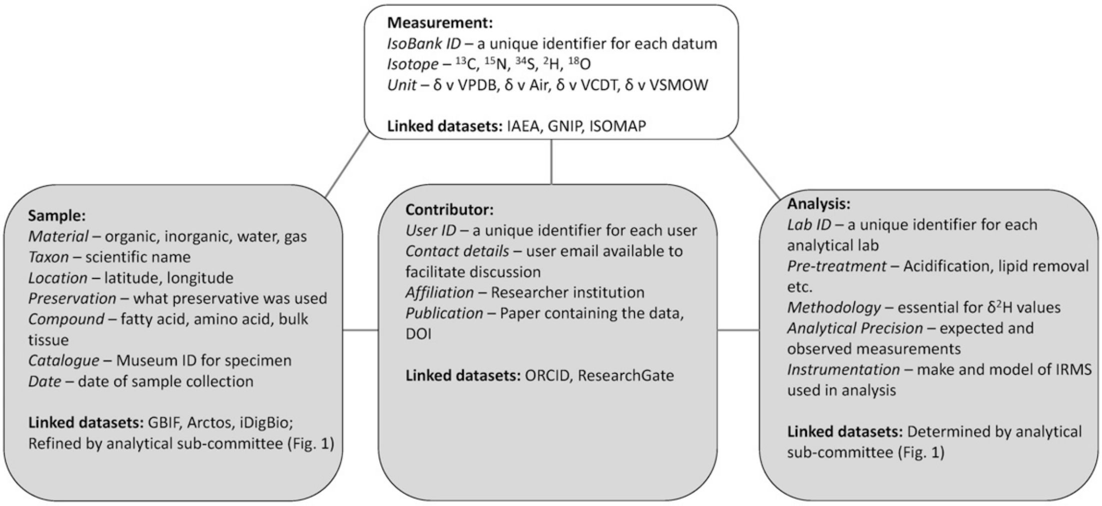

Fig. 2. A schematic of the proposed database structure, outlining how contributors and users would interface with samples, analyses, measurements, and datasets. GNIP, Global Network of Isotopes in Precipitation; IAEA, International Atomic Energy Association; VCDT, Vienna Canyon Diablo Troilite; VPDB, Vienna PDB; VSMOW, Vienna Standard Mean Oceanic Water.

are downloaded or used in subsequent publications. We also see value in assigning unique IDs to analytical laboratories for data uploads to provide an opportunity to compare and evaluate different methods, analytical standards, and precision among laboratories. Ultimately, profiles of individuals and laboratories with a range of optional metadata will better connect data generators to contributors to users, ultimately enhancing the use of stable isotope data.

To accommodate a wide range of researchers, each isotopic data record in IsoBank should be stored under a tiered framework. Initially, data will be stored in a subrepository (e.g., biogenic, inorganic, water), which will contain sufficient discipline-specific metadata to allow users to integrate data from IsoBank into discipline-specific or interdisciplinary analyses and to avoid handling irrelevant metadata terms (e.g., species taxonomy for water samples).

Sample metadata will fall under two categories: essential metadata, describing every data record in IsoBank, and discipline-specific metadata. To maximize the accessibility of IsoBank to data holders, the essential metadata should be kept to a minimum, and include latitude and longitude of sampling site, sample material, isotopes measured, and their values. Discipline-specific metadata will be developed by working groups during the initial phase of IsoBank.

Following the model established by the genomics community, the gold standard for accessions are data records that are tied directly to vouchered samples housed in permanent and accessible archives with data cross-linked to IsoBank, museum databases (e.g., Arctos), and data aggregators [e.g., iDigBio (15)]. If specimens are not curated in museums, users would be encouraged to provide sample storage location so that interested parties may contact them directly if they wish to conduct additional analyses.

Stable isotope data are produced in a wide range of research and commercial laboratories. Although the methods by which the majority of data, mostly bulk carbon (δ^{13}C) and nitrogen (δ^{15}N) stable isotope values, are generated is generally standardized, laboratories often use slightly different protocols and different laboratory reference materials to normalize data to internationally accepted scales (16). Other isotopes (e.g., δ^{2}H and δ^{18}O) have more fundamental issues associated with comparability measurements (17). To ensure data quality and user confidence in IsoBank, pertinent analytical information must be submitted for each data record. Therefore, mirroring the subdivisions of sample metadata, IsoBank should partition analytical fields into essential, recommended, and requested metadata. Such an approach will allow users with detailed analytical information to post it, but will not inhibit others who lack those details from depositing their data.

Essential metadata includes information, such as the specific isotope measured or the experimental error. In contrast, recommended and requested metadata may include sample pretreatment methods (e.g., lipid extraction, demineralization), analytical methods, instrumentation, or laboratory reference materials used to normalize data (18). The reliability and accuracy of data could subsequently be ranked from "moderately reliable" to "very reliable" by data managers at IsoBank, based on the level of analytical metadata provided.

Promoting Use

Given the successful model of GenBank, the direct application of isotopic data to pressing questions across diverse fields, and recent initiatives for data transparency and sharing, we believe high-quality data in IsoBank will be heavily used. Thus, our attention is focused primarily on procedures that will ensure deposition of high-quality and relevant data in IsoBank. To accomplish this, IsoBank should include features attractive to users as well as incentives to promote data-sharing.

First, we envision that IsoBank's graphical interface will enable users to easily navigate and query the database and rapidly upload and download data and associated metadata. We view IsoBank as a data repository and management system that features computational tools. However, the development of an application program interface would allow automated queries of the data and future integration with other datasets, a fundamental facet of Big Data analytics. Also, the structure of IsoBank's interface should be designed in such a way that it can also serve as a personal data management system to further incentivize use. This would encourage standardization between researchers and laboratories and would allow users to archive all their data under the IsoBank ontology, while also maintaining shared and private data archives.

The features of IsoBank that enable straightforward data uploads and analytical options would be paired with workshops and online assistance. To that end, IsoBank could follow the lead of other data repositories (e.g., ref. 15) and sponsor a series of workshops in the initial years at conferences, core isotope facilities, universities, and federal agencies to train potential users. Staff at IsoBank would also be available to respond to queries or problems that users encounter while using IsoBank. We would also seek collaborative opportunities with data-mining groups to harvest previously published stable isotope data from the peer-reviewed literature. In archiving these additional data, IsoBank could serve as a central online bibliography for publications that contain stable isotope data.

We believe that our shared vision for an IsoBank ... offers a viable and powerful framework to organize, consolidate, and broadly share stable isotope data across disciplines.

The development of IsoBank would create norms around data-sharing expectations among stable isotope scientists. To facilitate use of IsoBank, participants could place embargo periods on their datasets before public release. IsoBank staff would work closely with funding agencies to help incentivize its use for supported research (e.g., requiring the use of IsoBank in proposal data-management plans). This group would also work with the editorial boards of journals to ensure that deposition of data in IsoBank meets journal requirements for data accessibility before publication.

The value of inquiry-based approaches to education is now widely recognized (e.g., ref. 19) and motivates efforts to incorporate publicly available data into educational initiatives. Web-accessible data provide educators with excellent opportunities to build lessons that can engage students in original, data-driven exercises (20) and that promote the application of data to real-world problems, like climate change or disruption of biogeochemical cycles.

Such experiential and authentic lessons encompass the biological knowledge, analytical abilities, and computational skills needed by our next generation of scientists and policy makers to shape responses to these 21st century challenges. IsoBank would allow a diverse audience of students to directly access isotopic data for independent projects. We envision competitive IsoBank minigrants targeted to undergraduate students (e.g., National Science Foundation-Research Experiences for Undergraduates) who will conduct meta-analyses or quantitative reviews of isotopic data in their research projects.

Securing Funding

Given that stable isotopes are used by researchers globally, international opportunities should be pursued to fund IsoBank. To this end, we foresee IsoBank operating with independently funded mirror repositories as per GenBank (North America), EMBL (Europe), and INSDC (Japan). The large amount of start-up funding needed will likely require a collaboration between European and United States investigators. Applications to several European Union funding agencies, as well as similar agencies in the United States and Canada (e.g., National Science Foundation and National Sciences and Engineering Research Council), would facilitate a simultaneous start of both mirrors.

To ensure IsoBank's sustainability, we envision a long-term funding strategy that is part of governmental research infrastructure portfolios (e.g., National Institutes of Health support for GenBank), as well as funding from the community of stable isotope users. For example, revenue could be generated for IsoBank through a small fee-per-upload, whereby users pay a nominal amount to deposit their data. This model is already in use by some existing online repositories (e.g., Dryad) and represents regular income that should grow with the size and use level of the repository.

Imposing fees may potentially limit the use of IsoBank by researchers already facing constrained budgets. Thus, in the initial years of IsoBank, managers would need to ensure that data-deposit fees are manageable. In addition, IsoBank can engage directly with participating core laboratories to institute nominal surcharges per sample submitted (e.g., US$ 0.10–0.25 per sample). Given the hundreds of thousands of samples analyzed annually at core isotope facilities, this approach has the potential to generate sustained revenue to help offset the operation costs of IsoBank. By keeping fees low, the financial impact on researchers or laboratories would be limited. Finally, journal editors would need to ensure data deposition and availability in IsoBank by requiring authors to report data accession numbers in their manuscripts before publication, similar to current requirements for DNA data in GenBank.

As evidence of the immediate demand for an IsoBank, several websites are emerging (e.g., Neotoma, IsoMemo) to consolidate isotopic data within searchable databases. These have been launched within a variety of disciplines among international collaborators. We believe that our shared vision for an IsoBank—a single, comprehensive, and centralized repository managed by a team of experts and following a universally agreed ontology of metadata—offers a viable and powerful framework to organize, consolidate, and broadly share stable isotope data across disciplines. Such a repository would help to address the national initiative on data transparency, reinforce ongoing long-term and global data collection programs, and facilitate data integration as a tool to answer science's most challenging problems. We welcome a continued discussion to optimize the plan for IsoBank, but also see the need as extraordinary and encourage movement toward its rapid development and implementation.

Acknowledgments

We thank Brian Fry, Tamsin O'Connell, and Jim Ehleringer for constructive comments on an earlier daft of this manuscript; and the staff at the UNM Sevilleta Research Station for hosting the IsoBank Workshop. The IsoBank Workshop was funded with a grant through the National Science Foundation, Emerging Frontiers (NSF 1613214) and support from the Biodiversity Collections Network Research Coordinating Network (NSF 1441785). This article is dedicated to the memory of Scott Federhen.

1 Fry B (2006) *Stable Isotope Ecology* (Springer, New York).
2 West JB, Bowen GJ, Dawson TE, Tu KP (2010) *Isoscapes: Understanding Movement, Pattern, and Process on Earth Through Isotope Mapping* (Springer, The Netherlands).
3 Pauli JN, Steffan SA, Newsome SD (2015) It is time for IsoBank. *Bioscience* 65:229–230.
4 West JB, Bowen GJ, Cerling TE, Ehleringer JR (2006) Stable isotopes as one of nature's ecological recorders. *Trends Ecol Evol* 21(7):408–414.
5 Medina-Elizalde M, Rohling EJ (2012) Collapse of classic Maya civilization related to modest reduction in precipitation. *Science* 335(6071):956–959.
6 O'Brien DM (2015) Stable isotope ratios as biomarkers of diet for health research. *Annu Rev Nutr* 35:565–594.
7 Hall RO, Tank JL (2003) Ecosystem metabolism controls nitrogen uptake in streams in Grand Teton National Park, Wyoming. *Limnol Oceanogr* 48:1120–1128.
8 Boutton TW, Tyrrell HF, Patterson BW, Varga GA, Klein PD (1988) Carbon kinetics of milk formation in Holstein cows in late lactation. *J Anim Sci* 66(10):2636–2645.
9 Craine JM, et al. (2015) Convergence of soil nitrogen isotopes across global climate gradients. *Sci Rep* 5:8280.
10 Liu Z, et al. (2014) Paired oxygen isotope records reveal modern North American atmospheric dynamics during the Holocene. *Nat Commun* 5:3701.
11 Dumontier M, et al. (2014) The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. *J Biomed Semantics* 5(1):14.
12 Hinchliff CE, et al. (2015) Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proc Natl Acad Sci USA* 112(41):12764–12769.
13 Parr CS, et al. (2016) TraitBank: Practical semantics for organism attribute data. *Semant Web* 7:577–588.
14 Poelen JH, Simons JD, Mungall CJ (2014) Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. *Ecol Inform* 24:148–159.
15 Page LM, MacFadden BJ, Fortes JA, Soltis PS, Riccardi G (2015) Digitization of biodiversity collections reveals biggest data on biodiversity. *Bioscience* 65:841–842.
16 Ben-David M, Flaherty EA (2012) Stable isotopes in mammalian research: A beginner's guide. *J Mammal* 93:312–328.
17 Meier-Augenstein W, Hobson KA, Wassenaar LI (2013) Critique: Measuring hydrogen stable isotope abundance of proteins to infer origins of wildlife, food and people. *Bioanalysis* 5(7):751–767.
18 Jardine TD, Cunjak RA (2005) Analytical error in stable isotope ecology. *Oecologia* 144(4):528–533.
19 Feldman A, Chapman A, Vernaza-Hernandez V, Ozalp D, Alshehri F (2012) Inquiry-based science education as multiple outcome interdisciplinary research and learning (MOIRL). *Science Education International* 23:328–337.
20 Cook JA, et al. (2014) Natural history collections as emerging resources for innovative education in biology. *Bioscience* 64:725–734.